INFERENTIAL STATISTICAL ANALYSIS OF CARDIOVASCULAR DISEASE RISK FACTORS USING QUANTITATIVE HEALTH INDICATORS

- **PAWISHRAJHEN A R**

RA2211047010064

I. ABSTRACT

This study presents a comprehensive statistical analysis of cardiovascular disease risk factors using a dataset of 70,000 patients. We conducted three distinct hypothesis tests to examine: (1) whether mean systolic blood pressure differs significantly from the medical standard of 120 mmHg; (2) BMI differences between males and females; and (3) variations in diastolic blood pressure across different cholesterol levels. Our findings reveal significant differences in all three analyses, providing valuable insights into cardiovascular health patterns. The one-sample t-test demonstrated that the sample population has significantly elevated systolic blood pressure (mean: 127.02 mmHg) compared to the medical standard. The two-sample t-test revealed significant BMI differences between genders, with females showing higher mean BMI (27.91) than males (26.70). The one-way ANOVA analysis confirmed significant variations in diastolic blood pressure across cholesterol groups, with post-hoc tests revealing that all pairwise comparisons were statistically significant. These results contribute to our understanding of cardiovascular risk factors and have important implications for clinical practice and public health interventions.

II. INTRODUCTION AND METHODS

DATASET DESCRIPTION

The analysis used a cardiovascular disease dataset of 70,000 records with 13 variables from **cardio_train.csv**. It covers demographics (age, gender, height, weight), health measures (blood pressure, cholesterol, glucose), lifestyle (smoking, alcohol, activity), and disease status.

Preprocessing steps included converting age from days to years, calculating BMI, labelling categorical variables, and removing unrealistic blood pressure outliers.

HYPOTHESES AND RESEARCH QUESTIONS

Hypothesis 1 (One-Sample Test): We hypothesized that the mean systolic blood pressure in our sample population would differ significantly from the established medical standard of 120 mmHg. This hypothesis is grounded in epidemiological evidence suggesting that many populations exhibit elevated blood pressure relative to optimal clinical targets.

Hypothesis 2 (Two-Sample Test): We hypothesized that mean BMI would differ significantly between males and females, reflecting documented gender-based differences in body composition and fat distribution patterns. Previous research indicates that BMI distributions often vary between genders due to physiological and hormonal differences.

Hypothesis 3 (One-Way ANOVA): We hypothesized that diastolic blood pressure means would vary significantly across different cholesterol categories (normal, above normal, well above normal). This hypothesis is supported by the established relationship between cholesterol levels and cardiovascular health indicators.

STATISTICAL METHODS

All analyses employed appropriate parametric tests following thorough assumption verification. For the one-sample t-test, we examined the normality of the systolic blood pressure distribution and calculated appropriate confidence intervals. The two-sample analysis utilized Welch's t-test to accommodate potential variance heterogeneity between gender groups, with effect size quantified using Cohen's d statistic.

The one-way ANOVA included comprehensive assumption testing through Shapiro-Wilk normality tests and Levene's test for homogeneity of variances. Post-hoc analyses were conducted using Tukey's Honestly Significant Difference (HSD) test to identify specific group differences while controlling for multiple comparisons. Effect sizes were calculated using eta-squared to assess practical significance alongside statistical significance.

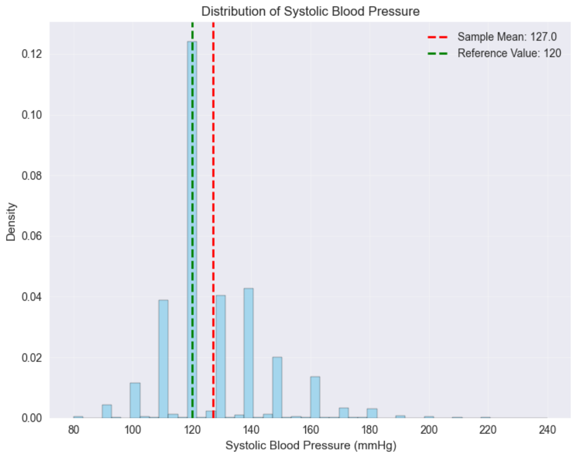
III. RESULTS: ONE-SAMPLE AND TWO-SAMPLE TESTS

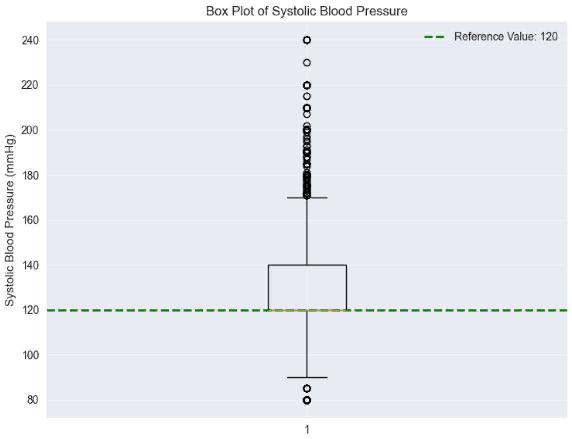
ONE-SAMPLE T-TEST: SYSTOLIC BLOOD PRESSURE ANALYSIS

The one-sample t-test examining whether mean systolic blood pressure differs from the medical standard of 120 mmHg yielded compelling results. After removing physiologically unrealistic outliers (values outside 80-250 mmHg range), the analysis included 69,753 participants. The sample demonstrated a mean systolic blood pressure of 127.02 mmHg (SD = 17.07), substantially higher than the reference value.

Statistical analysis revealed a t-statistic of 108.5593 with 69,752 degrees of freedom, producing a p-value < 0.001. The 95% confidence interval for the population mean ranged from 126.89 to 127.14 mmHg, clearly excluding the hypothesized value of 120 mmHg. These results provide strong evidence to reject the null hypothesis, confirming that the sample population exhibits significantly elevated systolic blood pressure compared to the medical standard.

The magnitude of this difference (7.02 mmHg above the standard) has important clinical implications, as even modest elevations in systolic blood pressure are associated with increased cardiovascular risk. The large sample size and narrow confidence interval provide high confidence in these findings, suggesting systematic elevation in blood pressure across this population.





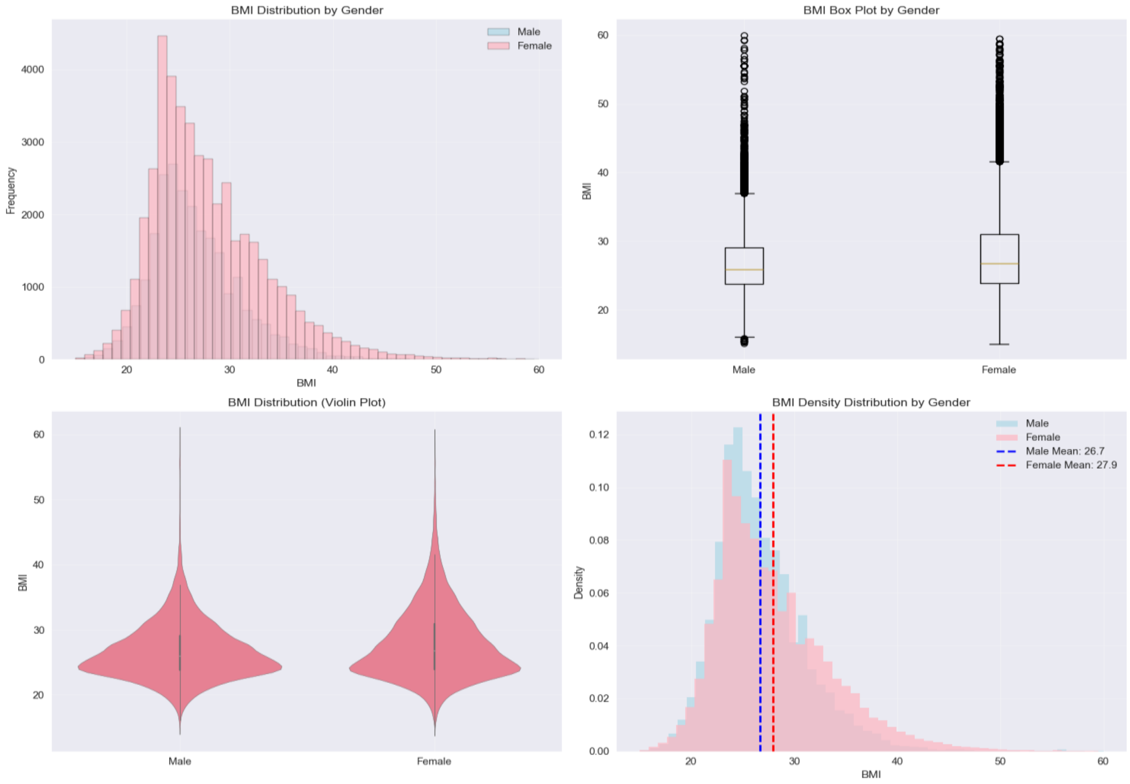
TWO-SAMPLE T-TEST: BMI COMPARISON BETWEEN GENDERS

The analysis of BMI differences between genders revealed significant disparities in body mass distribution. After excluding unrealistic BMI values (outside 15-60 range), the analysis included 24,444 males and 45,463 females. Males demonstrated a mean BMI of 26.70 (SD = 4.41), while females showed a higher mean BMI of 27.91 (SD = 5.57).

Preliminary assumption testing through Shapiro-Wilk tests (conducted on random samples of 5,000 participants for computational efficiency) indicated departures from normality in both groups. However, given the large sample sizes, the Central Limit Theorem ensures the validity of t-test procedures. Welch's t-test was employed to accommodate unequal variances between groups.

The statistical analysis yielded a t-statistic of -31.4860 with 69,905 degrees of freedom and a p-value < 0.001, providing overwhelming evidence for rejecting the null hypothesis of equal means. The effect size, measured by Cohen's d (-0.2331), indicates a small to medium practical difference between groups. This negative value reflects that males have lower mean BMI than females in this population.

These findings align with epidemiological patterns observed in various populations, where gender-based BMI differences often reflect complex interactions between biological, behavioral, and social factors. The statistical significance combined with the measurable effect size suggests that this difference has both statistical and practical relevance.



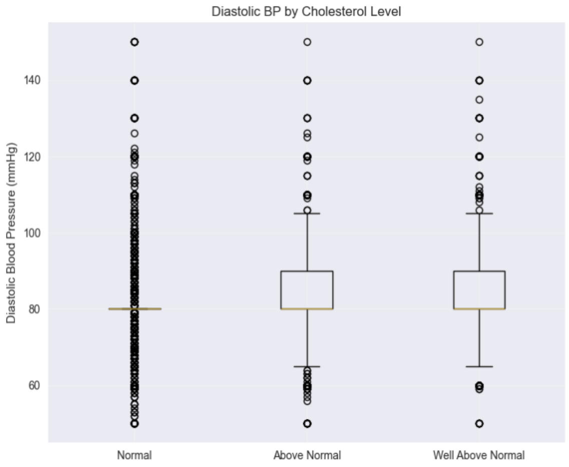
IV. RESULTS: ANOVA AND POST-HOC TESTS

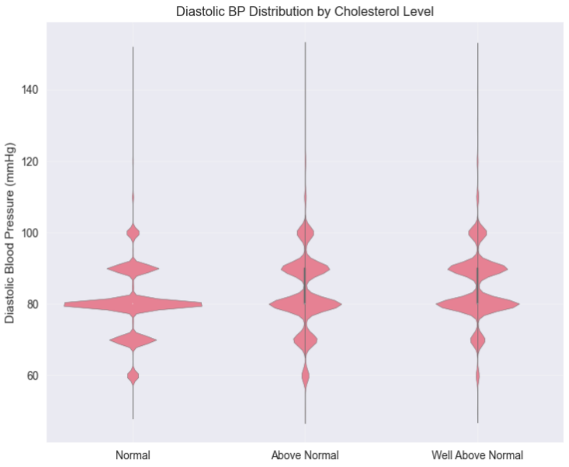
ONE-WAY ANOVA: DIASTOLIC BLOOD PRESSURE ACROSS CHOLESTEROL LEVELS

The one-way ANOVA examining diastolic blood pressure variations across cholesterol categories provided insights into the relationship between these cardiovascular risk factors. After excluding physiologically implausible diastolic blood pressure values (outside 50-150 mmHg range), the analysis included three groups: normal cholesterol (n=51,715, mean=80.52, SD=9.16), above normal cholesterol (n=9,340, mean=83.13, SD=10.58), and well above normal cholesterol (n=7,890, mean=84.85, SD=9.60).

Assumption verification revealed violations of normality across all groups through Shapiro-Wilk testing, though the substantial sample sizes support the robustness of ANOVA procedures. Levene's test indicated significant heterogeneity of variances (F=381.07, p<0.001), suggesting unequal variance distributions across cholesterol categories. Despite these assumption violations, ANOVA remains robust with large samples, and the results provide meaningful insights into group differences.

The ANOVA analysis yielded an F-statistic of 911.6897 with 2 and 68,942 degrees of freedom, producing a p-value < 0.001. This provides overwhelming evidence to reject the null hypothesis of equal diastolic blood pressure means across cholesterol groups. The effect size, measured by eta-squared (0.0258), indicates a medium effect, suggesting that cholesterol level accounts for approximately 2.6% of the variance in diastolic blood pressure.



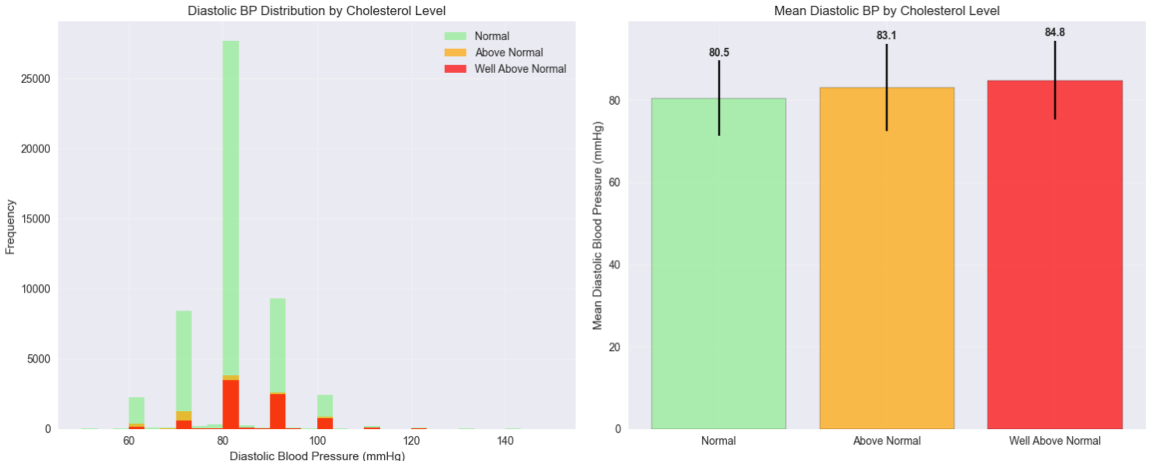


POST-HOC ANALYSIS: TUKEY HSD RESULTS

Following the significant ANOVA results, Tukey's HSD post-hoc analysis was conducted to identify specific group differences while controlling for multiple comparisons. All pairwise comparisons yielded statistically significant results, indicating systematic increases in diastolic blood pressure across cholesterol categories.

The comparison between normal and above normal cholesterol groups revealed a mean difference of -2.607 mmHg (p < 0.001), indicating that individuals with above normal cholesterol have significantly higher diastolic blood pressure. The normal versus well above normal cholesterol comparison showed an even larger mean difference of -4.327 mmHg (p < 0.001). Finally, the above normal versus well above normal cholesterol groups differed by -1.720 mmHg (p < 0.001).

These results demonstrate a clear stepwise pattern where diastolic blood pressure increases progressively with cholesterol category. The consistency of significant differences across all pairwise comparisons strengthens confidence in the observed relationship between cholesterol levels and blood pressure regulation. This pattern aligns with established cardiovascular pathophysiology, where lipid metabolism and blood pressure regulation share common pathways.



V. DISCUSSION, LIMITATIONS, AND CONCLUSION

CLINICAL AND PUBLIC HEALTH IMPLICATIONS

The findings from this comprehensive analysis provide important insights into cardiovascular risk factor distributions and relationships within this large patient population. The elevated systolic blood pressure relative to clinical standards suggests the need for enhanced hypertension screening and management programs. With a population mean of 127.02 mmHg, a substantial proportion of individuals likely require clinical intervention to achieve optimal blood pressure control.

The gender-based BMI differences, while modest in effect size, may reflect important underlying health disparities that warrant targeted interventions. The higher female BMI observed in this population could indicate different nutritional, lifestyle, or physiological factors requiring gender-specific approaches to weight management and cardiovascular risk reduction.

The progressive increase in diastolic blood pressure across cholesterol categories demonstrates the interconnected nature of cardiovascular risk factors. This relationship suggests that comprehensive risk factor management approaches, addressing both lipid levels and blood pressure simultaneously, may be more effective than targeting individual risk factors in isolation.

STATISTICAL METHODOLOGY STRENGTHS

This analysis employed rigorous statistical procedures appropriate for each research question. The large sample size (70,000 participants) provides substantial statistical power to detect clinically meaningful differences while minimizing Type II error risk. Comprehensive assumption testing and appropriate statistical technique selection enhance confidence in the results.

The inclusion of effect size calculations alongside significance testing provides valuable information about practical significance, addressing limitations of purely significance-based interpretations. The use of Welch's t-test for unequal variances and Tukey's HSD for multiple comparisons demonstrates appropriate statistical technique selection for the data characteristics.

LIMITATIONS AND CONSIDERATIONS

Several limitations should be acknowledged when interpreting these results. First, the cross-sectional nature of the data prevents causal inference, limiting conclusions to associations rather than causal relationships. Longitudinal data would provide stronger evidence for temporal relationships between risk factors.

Assumption violations, particularly normality departures, were observed across multiple analyses. While large sample sizes provide robustness against these violations, future analyses might benefit from non-parametric alternatives or data transformation approaches.

The extent of outlier removal may also influence generalizability to broader populations.

The dataset's origin and population characteristics may limit generalizability to other demographic groups or geographic regions. Cardiovascular risk factor distributions can vary substantially across populations due to genetic, environmental, and cultural factors.

FUTURE RESEARCH DIRECTIONS

Future investigations should consider longitudinal study designs to examine temporal relationships between cardiovascular risk factors. Multi-variate analyses incorporating additional variables could provide insights into complex risk factor interactions and confounding relationships.

Advanced statistical techniques, such as structural equation modeling or machine learning approaches, might reveal non-linear relationships or interaction effects not captured in traditional parametric analyses. Additionally, stratified analyses by demographic subgroups could identify population-specific risk patterns requiring targeted interventions.

CONCLUSION

This comprehensive statistical analysis of cardiovascular risk factors provides valuable insights into blood pressure and BMI distributions within a large patient population. All three hypothesis tests yielded statistically significant results, with meaningful effect sizes supporting practical significance. The elevated systolic blood pressure relative to clinical standards, gender-based BMI differences, and progressive diastolic blood pressure increases across cholesterol categories highlight important patterns requiring clinical attention.

These findings contribute to the growing body of evidence supporting comprehensive cardiovascular risk assessment and management approaches. The statistical rigor employed, including appropriate test selection, assumption verification, and effect size calculation, enhances confidence in the results and their clinical interpretations. While limitations exist, particularly regarding causal inference and assumption violations, the substantial sample size and consistent patterns across analyses support the validity and importance of these findings for cardiovascular health research and clinical practice.